Desarrollo de interfaces para la detección del habla sub-vocal

Interface developed for the detection of sub-vocal speech

Jenny Alejandra Gutiérrez Calderón
Ingeniera en Mecatrónica. Asistente de Investigación de la Universidad Militar Nueva Granada. Bogotá, Colombia. Contacto: gav@unimilitar.edu.co

Erika Nathalia Gama Melo
Ingeniera en Mecatrónica. Asistente de Investigación de la Universidad Militar Nueva Granada. Bogotá, Colombia. Contacto: gav@unimilitar.edu.co

Darío Amaya Hurtado
Ingeniero Electrónico. Doctor en Ingeniería Mecánica. Docente de la Universidad Militar Nueva Granada. Bogotá, Colombia. Contacto: dario.amaya@unimilitar.edu.co

Oscar Fernando Avilés Sánchez
Ingeniero Electrónico, Doctor en Ingeniería Mecánica. Docente de la Universidad Militar Nueva Granada. Bogotá, Colombia. Contacto: oscar.aviles@unimilitar.edu.co

Fecha de recepción: 6 de agosto de 2012
Clasificación del artículo: Revisión
Fecha de aceptación: 21 de mayo de 2013
Financiamiento: Universidad Militar Nueva Granada

***Palabras clave*:** habla sub-vocal, interfaz, reconocimiento de voz.

Key words: speech sub-vocal, interface, speech recognition.

RESUMEN
Por medio de este artículo se explorarán las técnicas más sobresalientes utilizadas actualmente para la detección del habla sub-vocal tanto en personas con parálisis cerebral como para aplicaciones comerciales (por ejemplo, permitir la comunicación en lugares ruidosos). Las metodologías expuestas se ocupan de adquirir y procesar las señales del habla desde diferentes niveles de su generación, de esta manera se presentan métodos que detectan y analizan señales desde que estas son producidas como impulsos neuronales en el cerebro, hasta que llegan al aparato fonador ubicado en la garganta, justo antes de ser pronunciadas. La calidad de la adquisición y procesamiento dependerá de varios factores que serán analizados en las siguientes secciones. La primera parte de este artículo constituye una breve explicación del proceso completo de generación de voz. Posteriormente, se exponen las técnicas de adquisición y análisis de las señales del habla sub-vocal, para finalmente incluir un análisis de las ventajas y desventajas que estas presentan

para su posible implementación en un dispositivo para la detección del habla sub-vocal o lenguaje silencioso. Los resultados de la investigación realizada demuestran cómo la implementación del micrófono NAM (Non-audible Murmur) es una de las alternativas que aporta mayores beneficios no solo para la adquisición y procesamiento de las señales, sino para la futura discriminación de los fonemas del idioma español.

ABSTRACT

This paper explores the most important techniques currently used to detect sub-vocal speech in people with cerebral palsy as well as for commercial purposes, (e.g. allow communication in very noisy places). The methodologies presented deal with speech-signal acquisition and processing. Signal detection and analysis methods are described throughout the whole speech process, from signal generation (as neural impulses in the brain) to the production sound in the vocal apparatus (located in the throat). Acquisition and processing quality depends on several factors that will be presented in various sections. A brief explanation to the whole voice generation process is provided in the first part of the article. Subsequently, sub-speech signal acquisition and analysis techniques are presented. Finally, a section about the advantages and disadvantages of the various techniques is presented in order to illustrate different implementations in a sub-vocal speech or silent speech detection device. The results from research indicate that Non-audible Murmur Microphone (NAM) is one of the choices that offer huge benefits, not only for signal acquisition and processing, but also for future Spanish language phoneme discrimination.

* * *

1. INTRODUCCIÓN

El habla constituye la forma más natural de comunicación entre las personas, de ahí el gran interés que tiene el desarrollo de sistemas informáticos capaces de procesarla y generarla de forma automática. El procesamiento del habla abarca un amplio abanico de métodos y técnicas que tienen una doble finalidad: por una parte, lograr que los ordenadores puedan comprender los mensajes pronunciados por los usuarios, y por otra, lograr que los usuarios puedan entender los mensajes generados por los ordenadores de forma oral [1]. Son diversas las aplicaciones del procesamiento del habla, entre las cuales se destacan los sistemas automatizados de información telefónica, programas de traducción entre idiomas, programas de dictado, entornos domóticos e inteligentes, sistemas de manejo oral de diversos aparatos, control oral de programas de ordenador, aplicaciones militares y seguridad, entre otras [2].

Sin embargo, la interfaz de control del habla basado en la acústica convencionalde las señales de voz aún experimenta muchas limitaciones, entre las que sobresalen dos. La primera limitación está en que las señales acústicas del habla son transmitidas por el aire y por tanto están propensas al ruido del ambiente. A pesar de los enormes esfuerzos aún no hay un sistema de procesamiento del hablaque proporcione buenos resultados en lugares ruidosos. La segunda radica en que las interfaces convencionales del habla se basan en el discurso en voz alta, que presenta el inconveniente de exponer al público comunicaciones confidenciales y perturbar a las personas que las escuchan[3].

Para superar estos problemas se propone que la captura y el procesamiento de la señal del habla se realicen antes de que el aire llegue al aparato fonador y así evitar que sea afectado desfavorablemente por las condiciones de ruido. La propuesta de lenguaje silencioso o habla sub-vocal, permitiría pronunciar cualquier discurso en silencio y

por tanto superar las limitaciones mencionadas. La información confidencial puede ser enviada en forma segura y el habla silenciosa no perturbaría o interferiría con el entorno[3].

El habla sub-vocal puede medirse (en principio) colocando sensores eléctricos en la lengua, cuerdas vocales y otras partes del aparato vocal [4]. Las señales biológicas surgen al leer o hablar con uno mismo, con o sin labios, o con el movimiento facial. Una persona que usa el sistema sub-vocal piensa en frases y habla consigo mismo en voz tan baja que no se puede oír, pero las cuerdas vocales y la lengua reciben las señales del habla transmitida desde el Sistema Nervioso Central (SNC).

Este método propuestoes lalectura directa de las señales del cerebro,cuyo enfoque evita la producción del habla por completo [5].Con el habla sub-vocal se tendrían aplicaciones en el área comercial (por ejemplo, con los teléfonos celulares silenciosos y la comunicación entre buceadores y astronautas) y se podría dar esperanzas a las personas con discapacidad en el habla afectada por una laringotomía o parálisis. Para captar estas señales neurológicas se debe entender "qué permite" la producción de la voz y "cómo lo hace".

2. LA VOZ

La producción de la voz inicia en la corteza cerebral. Existen interacciones complejas entre los centros del habla y la expresión musical y artística que establecen los comandos para la comunicación [6]. Este conjunto de instrucciones se trasmiten a los núcleos motores del tronco del encéfalo y la médula espinal, llevando la información a los músculos de la garganta, lo que permite la producción de la voz [7].

La producción de la voz se desencadena de un gran número de órdenes producidas por el sistema nervioso central, lo que genera una actividad coordinada de la musculatura laríngea, torácica, abdominal y las estructuras articuladoras y resonadoras [8]. El refinamiento de la actividad motora es regulado por el sistema extrapiramidal (corteza cerebral, cerebelo y ganglios basales) y el sistema nervioso autónomo [7]. Las neuronas del cerebro generan señales eléctricas muy pequeñas que pasan por el conjunto de nervios dentro de la columna vertebral, antes de desviarse a otras partes del cuerpo por medio del sistema nervioso. Una vez llegan al área correcta del cuerpo, se activan los músculos necesarios para completar una acción [9], [10].

Cuando una fibra muscular se activa por el sistema nervioso central, pequeñas corrientes eléctricas en forma de flujos de iones se generan. Estas corrientes eléctricas pasan a través del tejido del cuerpo, encontrando una resistencia que crea un campo eléctrico. La diferencia de potencial resultante se puede medir entre algunas regiones de la superficie corporal, es decir, en la piel. El registro de la actividad eléctrica muscular se hace mediante electrodos de superficie, esta señal eléctrica es amplificada y se obtiene a partir de la medición de la tensión a través del tiempo. La señal eléctrica muscular puede ser transmitida directamente a los dispositivos electrónicos para su posterior procesamiento [3]. Hasta la fecha, el procesamiento de estas señales se hace por medio de sistemas experimentales llamados Interfaz de Habla Silenciosa o Silent Speech Interface (SSI), que se basan en siete tipos de tecnologías.

3. INTERFAZ DE DETECCIÓN DE HABLA SUB-VOCAL

Cada uno de los siguientes apartados describe una tecnología diferente en cuanto a los métodos, ventajas, desventajas, el estado de desarrollo y rango de aplicaciones actuales de las interfaces de detección del habla sub-vocal.

3.1 Captura de movimientos usando sensores Electromagnetic Articulography (EMA)

Los sensores EMA pertenecen a la categoría de dispositivos de transducción que proporcionan datos sobre las trayectorias de los puntos articuladores

en dos dimensiones en el plano cartesiano. Son capaces de monitorear los movimientos sobre el plano medio sagital en la mayoría de las estructuras articulatorias que han sido enfocadas en los estudios de coarticulación como son los labios, la lengua, la mandíbula y el paladar. El principio de medición de los EMA empieza cuando se alterna un campo magnético generado por la bobina emisora. La fuerza de la señal inducida en un transductor (bobina receptora) es inversamente proporcional al cubo de la distancia entre el transmisor y el receptor [11].

Este método ha sido propuesto para la restauración quirúrgica de la voz después de una laringotomía. Algunos imanes son colocados en labios, dientes y lengua para generar el cambio magnético cuando el individuo dice palabras con la boca. Estos cambios son detectados por seis sensores con doble eje magnético implementados en unas gafas especiales, como se observa en lafigura 1.

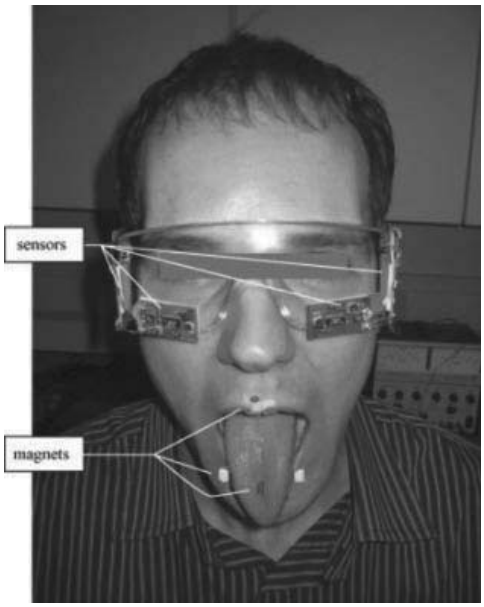

Figura 1. Figura de un sujeto que usa las gafas con sensores especiales y con imanes en la lengua, labios y dientes.
Fuente: tomada de [7].

La investigación que se describe se refiere a la identificación de palabras y/o fonemas de la boca del paciente. Se llevaron a cabo diez ensayos en los cuales se encontró una precisión del 97% en palabras y 94% en fonemas. Para la obtención de los resultados se utilizaron 12 sensores en total, adicionalmente se realizaron pruebas con sub-conjuntos de sensores para determinar la necesidad de utilizar todos en la obtención de los resultados. Con ocho sensores, la tasa de reconocimiento se redujo en un 93% para palabras, mientras que para los fonemas se redujo en un 87%. Con el subconjunto de cuatro sensores las tasas son del 84% y 58% respectivamente [12].

3.2 Caracterización en tiempo real del tracto vocal mediante ultrasonido (US)

Es una técnica de aprendizaje controlado por ultrasonido e imágenes ópticas, que posee dos métodos para su aplicación como una interfaz de lenguaje silencioso. El primero realiza un segmetalvocoder, un sistema construido sobre un diccionario audiovisual en donde se asocian imágenes con la acústica para cada clase de fonema. Las características visuales son extraídas de imágenes de ultrasonido de la lengua y los videos de los labios usando la codificación de imágenes basados en la técnica PCA (Análisis de Componentes Principales). Las observaciones visuales de cada clase fonética son modeladas por HMM (Hidden Markov Models) continuo. La propuesta de un SSI basado en ultrasonidos combina la etapa de reconocimiento fonético basado en HMM con la etapa de síntesis de las tarjetas fonéticas, este se basa en el registro de difonos que son buscados secuencialmente en el diccionario [13][14]. La figura 2 muestra el esquema para el SSI.

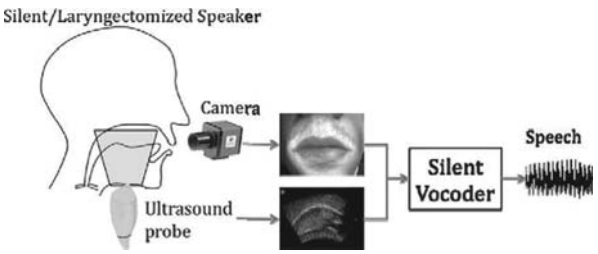

Figura 2. Esquema de ultrasonido para SSI
Fuente: tomada de [14].

El segundo método consta de una máquina con técnicas de aprendizaje usada para hacer coincidir la reconstrucción del contorno de la lengua con 30 fotogramas por segundo de las imágenes de ultrasonido del tracto vocal del hablante o ponente, a partir de una pista de audio sincronizada. El sintetizador del habla usa parámetros de aprendizaje y ruido, como una función de activación de muchas muestras de la frecuencia, que son características dominantes del audio original [15].

3.3 Transformación digital de la señal a partir de un micrófono Non-Audible Murmur (NAM)

El reconocimiento del soplo no audible NAM, por las siglas en inglés de Non-Audible Murmur, es una de las interfaces del lenguaje silencioso más prometedoras para la comunicación hombre-máquina. NAM es el término dado a los sonidos de baja amplitud generados por el flujo de aire en la laringe y su resonancia en el tracto vocal [41]. Este débil sonido del habla que es producido sin la vibración vocal, puede ser detectado usando especialmente un sensor: un micrófono de NAM [16]. Este dispositivo fue desarrollado por Nakajima, inspirado en un estereoscopio. El micrófono NAM fue originalmente desarrollado para detectar el murmullo extremadamente suave.

Un micrófono NAM comprende un micrófono condensador electret ECM (Electret Condenser Microphone), cubierto de un polímero suave, como la silicona o elastómero de uretano, que proporcionan una mejor impedancia y ayudan al contacto suave con el tejido del cuello. La sensibilidad del micrófono de acuerdo al material utilizado (silicona o elastómero de uretano) en 1kHz está entre – 41 y 58 dB. Como se observa en la figura 3, este se ubica en el cuello cerca de la oreja [17][18][19].

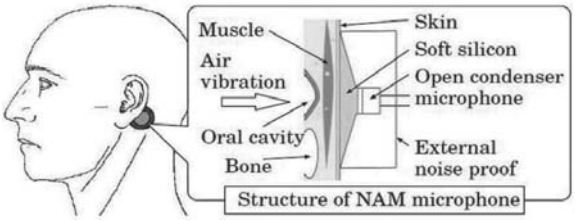

Figura 3. Posición y estructura del micrófono NAM.
Fuente: tomada de [17].

Las principales investigaciones referentes al reconocimiento de murmullo no audible (NAM) aplicado al lenguaje silencioso, se remontan a la introducción de un dispositivo especial capaz de detectar señales de este tipo a través de la piel, como es expuesto en [20]. La principal desventaja de este dispositivo es la pérdida de las altas frecuencias de la señal debido al medio de transmisión de las mismas, esto causa que las señales detectadas no sean lo suficientemente claras. En [42] se registra un estudio de la propagación del sonido desde el tracto vocal hasta la superficie del cuello con el objetivo de mejorar la claridad de la señales tipo NAM obtenidas.

El principal método para el reconocimiento de señales es el entrenamiento de un modelo acústico basado en el modelo oculto de Markov (Hidden Markov Model), como se explica en [43]. Basado en este mismo modelo, se presenta en [44] un nuevo método de reconocimiento NAM, el cual requiere solamente una pequeña cantidad de datos para el entrenamiento de HMM y está basado en adaptación supervisada e iteración adaptativa. Con el finde mejorar la claridad de las señales NAM, en [45] los autores reportan los resultados de un escaneo por imágenes de resonancia magnética del tracto vocal para ser aplicadas al estudio del mecanismo de producción del NAM y compararlo con el mecanismo de producción de habla normal.

Uno de los más recientes trabajos hechos sobre el análisis y reconocimiento de las señales tipo NAM es el registrado en [46]. Aquí los autores realizan el reconocimiento de los fonemas japoneses e incluyen el uso de señales provenientes del habla normal con

las señales tipo NAM para un mejor procesamiento. Esto debido a que las señales NAM muestreadas resultan débiles, por lo que requieren un proceso de amplificación antes de ser analizadas por herramientas de reconocimiento del habla. Este es un nuevo enfoque para métodos de reconocimiento de señales NAM que incluye la implementación de un modelo acústico y otro de lenguaje, así como la utilización de datos del habla normal transformados en datos NAM, como es posible observar en [47].

Para combatir el ruido causado entre otras cosas por el movimiento del hablante, en [52] se plantea el uso de un detector de señal estero junto a dos micrófonos NAMpara la supresión del ruido por medio de la estimación de fuentes ciegas separadas y de la substracción espectral que se realiza en cada uno de los canales. Un esquema de la utilización de esta tecnología para la detección del habla subvocal se muestra en la figura 4.

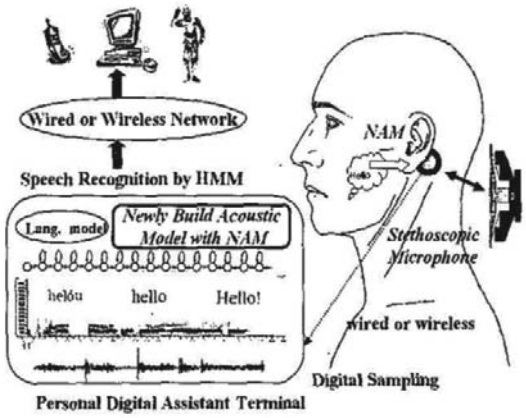

Figura 4. Concepto básico del método propuesto.
Fuente: tomada de [20].

Aunque existen ciertas desventajas referentes a la calidad del sonido al obtener señales NAM amplificadas, son muchas las aplicaciones que actualmente emplean el reconocimiento de señales NAM exitosamente:

- Desarrollo de sistemas de habla silenciosa. Aplicaciones concernientes a telefonía y reconocimiento del habla. El micrófono NAM permite establecer fácilmente una comunicación exitosa en situaciones donde la privacidad de la información es requerida o existe un entorno ambiental muy ruidoso. La amplificación de señales NAM produce resultados aceptables y las técnicas existentes de transformación del habla han sido aplicadas exitosamente para producir sonidos más naturales del habla [42].

- El dispositivo NAM es útil para hablantes con patologías de la voz debido a trastornos de la laringe, o individuos con dificultades para hablar debido a su edad [48].

- La simple amplificación del habla tipo NAM es benéfica para aplicaciones en conversación, lectura y llamadas telefónicas. En [49] se presenta un estudio de las técnicas implementadas para la captura del murmullo audible, en el marco del habla silenciosa aplicado a las comunicaciones telefónicas.

3.4 Análisis de la actividad de la glotis con electromagnetismo o sensores de vibración

El principio básico de este estudio es obtener formas de ondas de la glotis, que pueden ser usadas para eliminar el ruido en correlación con la señal acústica obtenida de un micrófono estándar para hablar de cerca. Son varios los sensores que se han desarrollado bajo el principio electromagnético y basan su funcionamiento en la propagación de la energía en forma de onda, presentando gran simpleza en su implementación, debido principalmente a su principio físico de funcionamiento [26]. Basados en este principiose encuentran los sensores nonacoustic, que proporcionan mediciones en función de la excitación de la glotis, es decir, movimiento del tracto articulador vocal, que son acústicas prácticamente inmunes a disturbios y pueden suplementar la forma de la onda acústica del habla [21]. Los sensores de movimiento electromagnético GEMS, por sus siglas en inglés (General Electromagnetic Motion System), son aquellos

que miden el movimiento del tejido durante el habla sonora cuando esta envuelve de vibraciones a las cuerdas vocales. Son sensores radio frecuencia (RF) que se colocan directamente en contacto con la piel y fueron desarrollados por Aliph Corporation. Miden la vibración de la pared de la tráquea durante el discurso oral, mientras se emite una señal electromagnética de 2GHz que penetra en la piely se refleja en la anatomíade producción del habla, que está compuesta por la traquea, las cuerdas vocales o la pared del tracto vocal [27], como se muestra en la figura 5.

Las señales recogidas del dispositivo GEM dependerán del movimiento del tejido de la anatomía de la producción del habla, que es relativamente inmune a la degradación debido a fuentes externas deruido acústico [22]. Los TERC (Tuned Electromagnetic Resonatic Collar) son sensores que miden los cambios de la capacitancia eléctrica de la glotis, basados en imágenes de resonancia magnética.Estos sensores no requieren de una posición precisa, ya que el sensor está diseñado para detectar pequeñas perturbaciones dieléctricas características del tejido del cuello, que resultan del ciclo de la glotis durante el habla. El funcionamiento de esta técnica se basa en uno o más capacitores que son situados alrededor del tejido del cuello mediante la colocación de dos o más placas conductoras sobre el collar. Aunque no se requiere que esté en contacto con el cuello, se puede usar así por conveniencia [28][29].

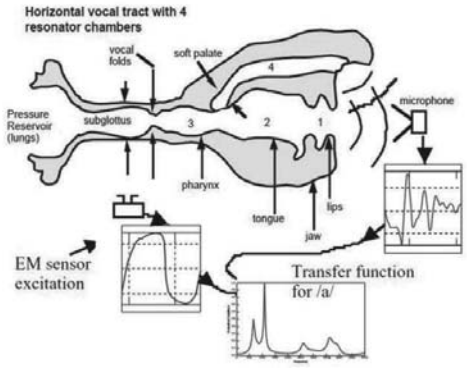

Figura 5. Ubicación de los sensores GEM, la excitación correspondiente y funciones acústicas y función de transferencia resultante.

Fuente: tomada de [27].

Igualmente, se cuenta con los sensores basados en el principio de vibraciones, entre ellos el P-mic (Physiological microphone), un sensor piezoeléctrico con un gel pack para el contacto con la piel humana. Este fue desarrollado por los laboratorios de investigación del ejército para medir el proceso fisiológico de la frecuencia cardiaca y respiratoria, y desde entonces ha sido utilizado como sensor [23], [24]. Los P-mic se localizan en la garganta, proporcionando una buena atenuación de ruido (30 dB), con una buena información de excitación localizada debajo de la glotis. La señal del P-mic tiende a ser pasa-bajo, con un desempeño significativo por encima de 2, 1kHz [25].

Otros sensores menos utilizados para adquirir señales del habla sub-vocal son: el electro glotto graph (EGG), una herramienta de investigación estándar que fue diseñada para detectar cambios en la impedancia eléctrica a través de la garganta durante el discurso sonoro. Se compone de dos electrodos con una capa de oro, ubicados a cada lado de la laringe por medio de un collar, con un potencial aplicado.Cuando las cuerdas vocales se cierran, la impedancia eléctrica disminuye, mientras que cuando están abiertas,un valor más alto se produce. Las vibraciones en la glotis inducen una señal de aproximadamente 1V RMS en una frecuencia de 3,2MHz, que facilita la vibración. Una desventaja de esta técnica es la sensibilidad de la colocación exacta de los electrodos [21].

3.5 Electromiografía de superficie (sEMG) basado en el reconocimiento del habla

La utilización de esta tecnología consiste en registrar la actividad eléctrica del músculo, la cual es capturada por electrodos de superficie (es decir, no implantados), produciendo señales acordes con la vibración o movimiento del aparato fonador. Las señales eléctricas son producidas por la fibra muscular, que es activada por el sistema nervioso central, lo que genera una pequeña descarga de corriente eléctrica en forma de flujos de iones. Estas corrientes eléctricas se muevena través de los teji-

dos del cuerpo,cuya resistencia crea diferencias de potencial que se pueden medir entre las diferentes regiones de la superficie corporal, por ejemplo, en la piel. Después de ser adquiridas las señales, se realiza un proceso de amplificación para tener señales que puedan ser utilizadas en la reproducción de la voz [29] - [32]. La ubicación de los electrodos se muestra en la figura 6.

Figura 6. Posicionamiento de los electrodos EMG.
Fuente: tomada de [31].

La aplicabilidad del reconocimiento del habla basado en EMG en ambientes acústicos hostiles ha sido investigada por la NASA. Su investigador principal, Chuck Jorgensen, ha logrado un 74% de precisión sobre la clasificación de 15 palabras, en un sistema de tiempo real que fue aplicado a sujetos que están expuestos a 95dB de nivel de ruido.

Hay también investigaciones interesantes que se trabajan en Estados Unidos, donde Ambient Corporation ha desarrollado un sistema en el cual las entradas son las señales de superficie EMG de uno o más electrodos situados por encima de la laringe. El propósito de este sistema es transformar las señales obtenidas por medio de EMG en voz, que es sintetizada a partir de las imágenes captadas por una camara, a las cuales se les asigna una señal tipo EMG. El funcionamiento de evaluación de este sistema fue probado para oírcinco vocales japonesas, las cuales reconoció en un rango del 76,8 %. El sistema consiste en tres módulos (figura 7): el primero es la entrada que corresponde a la imagen de la pronunciación oral obtenida de una cámara, la cual se usó en el contorno de los labios; el segundo es la estimación que se hace en función del área del tracto vocal y se extrae desde las características consideradas para la condición del órgano articulatorio; y el tercero, las señales del habla que se sintetizan con un filtro sintetizador PARCOR [34][35].

Figura 7. Sistema propuesto para la imagen de entrada del micrófono.
Fuente: tomada de [34].

Otros trabajos realizados basados en el sistema EMG, proponen un componente independiente de análisis (Independet Component Analysis[ICA]) para la extracción de las características y una clasificación por medio de Rede Neuronal. Usan fonemas de las vocales de una base de datos con un éxito de 93,99%. Este sistema consta de tres fases: adquisición, aprendizaje y clasificación. En la fase de adquisición, se usa DAS (Data Acquisition System), compuesto por una tarjeta de adquisición, sensores (electrodos de superficie) y scripts en Matlab. En la fase de aprendizaje, con ICA se aprenden las funciones base de los datos de entrada, y en la fase de clasificación se usa la red neuronal. La figura 8 muestra un diagrama de bloques propuesto a partir de este sistema [36].

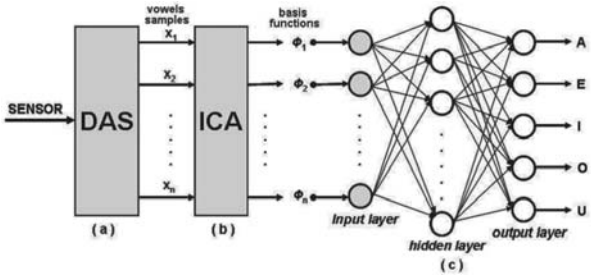

Figura 8. Propuesta del sistema de reconocimiento de habla subvocal. (a) Adquisición, (b) Extracción de característica, (c) Clasificación.
Fuente: tomada de [36].

En [51] se describen diferentes estrategias de adquisición y procesamiento de las señales de habla obtenidas por medio de sMEG, que ayudan a superar las limitaciones propias de este método, como la acústica del reconocimiento del habla, y plantean la posibilidad de aplicar este método en sistemas de comunicación y aplicaciones para el control de dispositivos.

3.6 Interpretación de las señales por sensores de electroencefalografía (EEG)

Además de las aplicaciones clínicas conocidas, la EEG ha demostrado ser útil para una multitud de nuevos métodos decomunicación, a través de la captura y procesamiento de las señales que se producen en el cerebro al momento de pensarse en la pronunciación de una palabra [33][37][50]. Es una técnica invasiva alternativa para la aplicación de dispositivos de interface cerebro-computador (BCI). Este sistema requiere del implante de un electrodo especial en la capa externa del neocortex. El archivo de las señales es transmitido al receptor más cercano y procesado para ser controlado por el cursor en el monitor de un computador que está al frente del paciente [38]. Los tres elementos que se analizan sobre las señales obtenidas del motor cortical humano son: 1. Detección de la señal: donde se identifica el potencial de acción desde grabaciones electricas extracelulares; 2. La tasa de estimación: que cuantifica los patrones de activación neuronal; y 3. La identificación de la señal se realiza mediante un mapeo de actividad neuronal para un uso determinado [39]. El esquema propuesto se muestra en la figura 9.

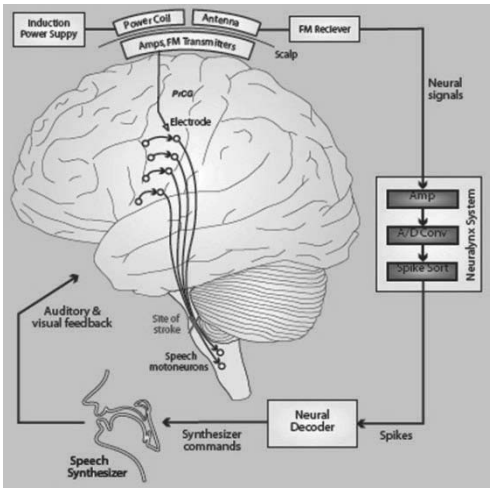

Figura 9. Sistema propuesto utilizando micro-electrodos intracordiales.
Fuente: tomada de [39].

Este método permite restaurarla comunicación oral de las personas paralizadas, o restablecer la comunicación por escrito, a través del desarrollo de un sistema que facilita el control de un cursor de un mouse, el cual se puede implementar en un teclado virtual. Existen varios factores a tener en cuenta cuando se quiere utilizar el método de microelectrodo intracortical para SSI, entre los cuales están la elección de los electrodos y la modalidad de decodificación [40]. El éxito del SSI intracortical requiere de electrodos que puedan ser implantados en los humanos, estos deben ser durables o proporcionar observaciones consistentes de las señales neurales.

La modalidad de decodificación es importante para el desarrollo de la prótesis neural. Con"modalidad"se hace referencia a la naturaleza de la señal decodificada o interpretada a partir de la observación de la actividad neural. Durante los últimos años, los electrodos intracordiales han sido utilizados en pacientes voluntarios con parálisis cerebral severa.

4. ANÁLISIS DE RESULTADOS

Las expuestas anteriormente son las tecnologías disponibles para reconocimiento de habla silenciosa

o habla sub-vocal que se han venido desarrollando e implementando en los últimos años. Estas técnicas de SSI cubren la extracción de información proveniente de todos los estados de la producción de la voz, desde la intención de habla hasta el efecto de articulación. La figura 10 resume los SSI expuestos anteriormente de acuerdo a las fases en las cuales se producen las señales analizadas.

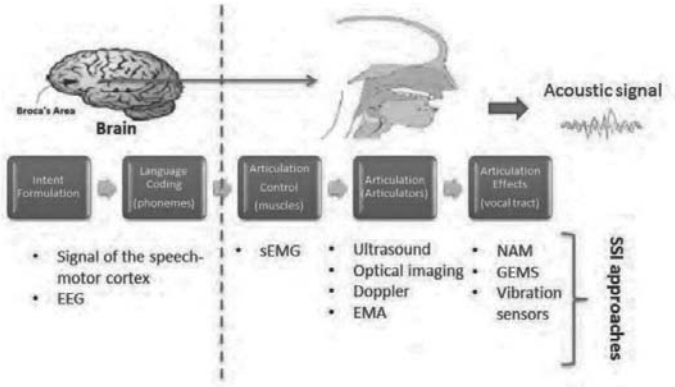

Figura 10. Modelo de las fases de producción del habla con sus correspondientes SSI.
Fuente: tomada de [48].

Para entrar a analizar cada una de estas interfaces, es necesario definir los puntos más críticos y que necesitan mayor avance e investigación para lograr un sistema cómodo, eficaz y con un margen de error casi despreciable en la interpretación de los fonemas. La tabla 1 muestra algunos de los retos que actualmente enfrentan las SSI y el estado en el que se encuentra cada una de ellas, comparada con las demás. Primero, se evalúa el grado de afectación que en el comportamiento de la interfaz pueden representar los cambios en la posición de los sensores que se generan durante la adquisición y síntesis de las señales. Cuando se trabaja con sensores ópticos o de ultrasonido, por ejemplo, cada movimiento del hablante durante la toma de datos puede significar un cambio en el marco de referencia de la imagen, lo que hace que los resultados obtenidos sean "sesión-dependientes", es decir, cada vez que se implemente la interfaz, deben establecerse nuevos parámetros de medición relativos a cada sesión.

Tabla 1. Comparación de algunas características de las seis interfaces.

ISS/ **Característica**	**Posición de los sensores**	**Independencia del hablante y comodidad**	**Permite su uso comercial**	**Permite su uso en medicina**
EMA	Sensible a variación	No es cómodo, no brinda independencia	No	Sí
Optimo y de ultrasonido	Sensible a cambio de marco de referencia de la imagen	No muy cómodo, no brinda independencia	Con limitaciones	Sí
NAM	Es necesario encontrar el punto más óptimo	Cercano a la independencia y cómodo	Sí	No
Sensores de vibración y EM	Es necesario encontrar el punto más óptimo	Cierto grado de independencia, ciertamente cómodo	Sí	No
EMG	Poca sensibilidad a variación	Poca independencia, relativamente cómodo	Sí	Sí
EEG	Sensible a variación	No permite independencia, no es cómodo	No	Sí

Fuente: elaboración propia.

Posteriormente se evalúa la comodidad del hablante cuando está usando la interfaz y la independencia que esta brinda para realizar tareas cotidianas. En esta parte también está relacionado el hecho de que la interfaz sea o no invasiva para el hablante. Finalmente, se evalúa el campo de aplicación de cada uno de ellos, para uso comercial o para uso médico, debido a que estas son las aplicaciones más comunes de este tipo de interfaces. El uso comercial se refiere a si es o será en un futuro próximo posible utilizar la interfaz en, por ejemplo,comunicaciones móviles, en sistemas de seguridad de la información, etc. Esta aplicación requiere, entre muchas más cosas, que la interfaz no se vea afectada por el ruido a veces excesivo del ambiente. La rehabilitación médica está directamente relacionada con la posibilidad de utilizar la interfaz en pacientes que han sido sometidos a una laringectomíau otra patología similar, para mejorar su calidad de vida.

Partiendo de estas premisas y observando con detalle las características presentadas por cada interfaz, consideramos que la detección del murmuro no audible (NAM), por su grado de evolución y la independencia que le permite al hablante, es una de las técnicas más prometedoras para su aplicación comercial en una interfaz de control hombre-máquina y para la detección de los fonemas del idioma español, ya que hasta el momento los principales desarrollos que se han hecho con este sistema son en la detección de fonemas japoneses.

5. CONCLUSIONES

A lo largo de este artículo se han revisado las seis tecnologías con mayor capacidad para el reconocimiento de las señales provenientes del habla subvocal, demostrando las ventajas y falencias de cada una de ellas. Algunas muestran mayor desarrollo que otras, como la electromiografía de superficie y el murmuro no audible (NAM). Es posible notar los desafíos actuales y las futuras aplicaciones para los sistemas de adquisición y procesamiento de señales provenientes del proceso de habla, entre las cuales se encuentran los próximos campos de acción en comunicaciones, seguridad, y control de dispositivos.

Entre los obstáculos a superar está la supresión de ruido y la implementación de mejoras en cuanto al rango de reconocimiento, la robustez, y la interfaz con el usuario; esta última se refiere a que el sistema implementado no afecte las actividades del hablante y pueda funcionar bajo condiciones adversas y diferentes tipos de entorno (ambientes ruidosos). Asimismo, es fácil observar que la efectividad y rango de acierto en la identificación de fonemas y palabras aumenta a medida que se adquieren señales más próximas a los impulsos generados desde el celebro, por lo que se hallan mayores errores cuando se implementan métodos como el basado en sensores tipo EMA. Sin embargo, tales métodos representan mayor desafío en la etapa de adquisición y procesamiento de las señales.

Este trabajo demuestra las múltiples alternativas existentes para obtener un sistema de comunicación silencioso que puede estar supeditado a las características propias del usuario o del entorno en el cual se vaya a trabajar. También demuestra el amplio campo de investigación que aún hay en esta temática y las posibles aplicaciones que pueden derivar del constante avance de las interfaces de lenguaje silencioso.

REFERENCIAS

[1] R. López-Cózar, and M. Araki. *Spoken, Multilingual and Multimodal Dialogue Systems: Development and Assesment*. Inglaterra: John Wiley & Sons, pp. 1851-189, 2005.

[2] V. Ceballo. *Manual de técnicas y modificación de conducta*. Madrid: Editorial Siglo XXI de España,2008.

[3] T. Schultz and M. Wand, "Modeling coarticulation in EMG-based continuous speech recognition",Speech Communication Journal, vol. 52, pp. 341-353, Dic.2009.

[4] NASA. *La NASA desarrolla el sistema para Automatizar el "Habla Sub-Vocal"*. [en línea]. [abril 2004]. Disponible: http://www.nasa.gov/centers/ames/spanish / news/releases/2004/04_18AR_span_prt.htm [agosto 2011]

[5] J.R. Wolpaw, N. Birbaumer, W.J. Heetdrecks, D. McFarland, P.H. Peckham, G. Schalk, E. Donchin, L.A. Quatrano, C. Robinson, and T.M Vaughan, "Brain-computer interface technology: a review of the first international meeting", IEEE Transactions on Rehabilitation Engineering, vol. 8, no. 2, pp. 164-173, Jun.2000.

[6] Pontificia Universidad Javeriana. Libro de Neurobioquímica: Cerebro. [en línea]. [agosto 2011]. Disponible: http://www.javeriana.edu.co /Facultades/Ciencias/neurobioquimica/libros/neurobioquimica/CEREBRO.htm [agosto 2011]

[7] R. Sataloff, Y. Heman-Ackah, and M. Hawksha. "Clinical anatomy and physiology of the voice", Otolaryngol Clinics of North America, vol. 40, no. 5, pp. 909 - 929, Oct.2007.

[8] A. Rubin, and R. Sataloff. "Vocal fold paresis and paralysis: what the thyroid surgeon should know",Surgical Oncology Clinics of North America, vol.17, no. 1, pp. 175-196,Ene.2008.

[9] Medicina y Farmacología,Circunvolución Frontal Ascendente, [en línea] [marzo 2010], Disponible: http://medicinafarmacologia.blogspot.com/2010/03/circunvalacion-frontal-ascendente.html

[10] E.B. Goildstein,Sensación y Percepción. Madrid: Editorial Thomson, pp. 19-23,2006.

[11] P. Hoole, N. Nguyen, W. Hardcastle, and N. Hewlett. "Coarticulation: Theory, Data and Techniques". New York: Cambridge University Press, vol. 1, p. 260, 1999.

[12] M. Fagan, S. Ell, J. Gilbert, E. Sarrazin,and P.M. Chapman,"Development of a (silent) speech recognition system for patients following laryngectomy", Medical Engineering & Physics, vol. 30, no. 4, pp. 419-425, Mayo 2008.

[13] T. Hueber, E.L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stono. "Development of a silent speech interface driven by ultrasound and optical images or tongue and lips",Speech Communication,vol.52,no. 4, pp. 288-300,Mar.2010.

[14] B. Denby, Y.Oussar, G. Dreyfus,and M. Stone."Prospects for a Silent Speech Interface Using Ultrasound Imaging".IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP. Toulouse, Francia,pp.365-368,Jul.2006.

[15] B. Denby and M. Stone."Speech Systhesis from real time ultrasound images of the Tongue",IEEE International Conference on Acoustics, Speech, and Signal Processing,vol.1,no. 1, pp. 685 - 688, May. 2004.

[16] Y. Nakajima, and K. Shikano, "Methods of fitting a non-audible murmur microphone for daily use and development of urethane elastomer duplex structure type non-audible murmur microphone",Journal

of the Acoustical of America,vol.120,no. 5, p.3330,Dic.2006.

[17] T. Toda, K. Nakamura, T. Nagai, T. Kaino, Y.Nakajima and K. Shikano."Technologies for processing Body-Conducted Speech Detected with Non-Audible Murmur Microphone", 10th Annual Conference of the International Speech Communication Association Interspeech,pp. 632-635,Sep.2009.

[18] S. Shimizu, M. Otani, and T. Hirahara. "Frecuency characteristics of several non.audible murmur (NAM) microphones",Acoustical Science and technology the Acoustical Society of Japan,vol.30,no. 2, pp. 139-142,Dic.2009.

[19] Y. Nakajima, H. Kashioka, N. Campbell, and K. Shikano, "Non-Audible Murmur (NAM) Recognition",Journal IEICE Trans. Information and Systems, vol. E89-D, no. 1, pp. 1-8, Ene. 2006.

[20] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell."Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin",IEEE ICASSP, vol. 5, pp. 708-711, May. 2003.

[21] T.Quatieri,D.Messing,K.Brady,W.Campbell, J. Campbell, M. Brandstein, C. Weinstein, J. Tardelli, and P. Gatewood,"Exploiting nonacoustic sensors for speech enhancement", IEEE Transsactions on Audio Speech Languaje Processing,vol. 14, no. 2, pp. 533-544,Mar.2006.

[22] G.C. Burnett, J. F. Holzrichter, T. J. Gable, and L. C. Ng, "The use of glottal electromagnetic micropower sensors (GEMS) in determining a voiced excitation function",Journal of the Acoustical Society of America,vol. 106, no. 4, pp. 2183-2184,Nov.1999.

[23] M. V. Scanlon, "Acoustic Sensor for Health Status Monitoring", Proceedings of IRIS Acoustic and Seismic Sensing,vol. 2, pp. 205-222,1998.

[24] J.D. Bass, M. V. Scanlon, T. K. Mills, and J. J. Morgan, "Getting Two Birds with One Phone: An Acoustic Sensor for Both Speech Recognition and Medical Monitoring",Journal of the Acoustic Society of America,vol. 106,no. 4, pp. 2180,Nov. 1999.

[25] K. Brandy, T.F. Quatieri, J.P. Campbell, W.M. Campbell, M. Brandstein, and C.J. Weinstein, "MultisensorMELPe using parameter substitution",*IEEE International Conference on Acoustics, Speech, and Signal Processin*,vol. 1,no. 1, pp. 477-480, May. 2004.

[26] M. Ratner, and D. Ratner. Nanotecnology: A gentle introduction the next big idea, New Jersey: Estados Unidos: Editorial Pearson Education,no. 1,pp. 100-102, 2001.

[27] G. C. Burnett, J. F. Holzrichter, T. J. Gable, and L. C. Ng, "Denoising of Human Speech using Combined Acoustic and EM sensor Signal Processing". International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Istanbul, Turquía, Jun. 1999.

[28] D.R. Brown, K. Keenaghan, S. Desimini, "Measuring glottal activity during voiced speech using a tuned electromagnetic resonating collar sensor", Measurement Science and Technology,vol. 16,no. 1,pp. 2381-2390, Jun. 2004.

[29] G. Bogdanov and R.Ludwig. "Coupled microstrip line transverse electromagnetic resonator model for high-field magnetic resonance imaging",Magnetic Resonance in Medicine,vol.47,no. 3,pp. 579-593, Mar. 2002.

[30] C. Jorgensen, and K. Binsted, "Web browser control using EMG based sub vocal speech recognition", 38th Annual Hawaii International Conference on System Sciences IEEE,vol. 38,pp.294c-294c, Ene. 2005.

[31] L. Maier-Hein, F. Metze, T. Schultz, and A.Waibel. "Session independent non-audible speech recognition using surface electromyography", IEEE Workshop on Automatic Speech Recognition and Understanding, pp.331-336, Ene. 2006.

[32] T. Hasegawa, and K. Ohtani,"Oral image to voice converter, image input microphone",Communications on the Move IEEE ICCS/ISITA,vol.2,no. 1,pp. 617-620, Ago. 2002.

[33] A. Porbadnigk, M.Wester, J.Calliess, and T. Schultz."EEG-based speech recognition - impact of temporal effects",Biosignals, pp.376 - 381, 2009.

[34] T. Hasegawa, K. Ohtani, and K. Oral."Image to voice converter, image input microphone".Proc. IEEE ICCS/ISITA 1992 Singapore, vol. 20, no. 1, pp. 617-620, 1992.

[35] K. Otani,and T. Hasegawa. "TeheImagen Input-Microphone- A new nonacoustic Speech Comunitacion System by Media Conversion from Oral Motion Images to Speech",IEEE Journal on Selected Areas in Communications,vol. 13,no. 1,pp. 42 – 48,Ene. 1995.

[36] J.A. Mendes, R.R. Robson, S. Labidi, and A.K. Barros,"Sub-vocal Speech Recognition Base on EMG signal Using Independent Component Analysis and Neural Network MLP", Congress on Image and Signal Processing,pp. 221-224, May. 2008.

[37] M. Wester, and T.Schultz, "Unspoken speech - speech recognition based on electroencephalography",Tesis de Maestría, Universita¨t Karlsruhe, Karlsruhe, Germany, 2006.

[38] P. R. Kennedy, R. A. E. Bakay, M. M. Moore, K. Adams, and J. Goldwaithe, "Direct Control of a Computer from the Human CentralNervous System",IEEE Transactions on Rehabilitation Engineering,pp. 198-202,Jun.2002.

[39] J.S. Brumberg, A. Nieto-Castanon, F.H. Guenther, J.L. Bartels, E.J. Wright, S.A. Siebert, D.S Andreasen, and P.R. Kennedy. "Methods for construction of a long-term human brain machine interface with the Neurotrophic Electrode",Neuroscience Meeting Planner, pp.779- 784, 2008.

[40] X. Huang, A.Acero, and H.Hon, Spoken Language Processing: A Guide to Theory, Algorithm and System Development,Washington, D.C.: Prentice Hall, 2001.

[41] B. Denby, T. Schultz, K. Honda, T. Hueber, J.M. Gilbert, and J.S. Brumberg. Silent speech interfaces,Speech Communication, vol. 52, no. 4, pp. 270-287, 2010.

[42] M. Otani, T.Hirahara, and S.Adachi, Numerical simulation of attenuation characteristics of soft-tissue conducted sound originated from vocal tract,In 19° International congress on acoustics, Madrid,2007.

[43] M.Clements, andS. Lim, Hidden Markov Model speech recognition based on Kalman filtering.School of Electrical Engineering, Georgia. IEEE. 0987.

[44] P.Heracleous,Y.Nakajima,A.Lee, H. Saruwatariand K.Shikano,Accurte Hidden Markov Models for Non-audible Murmur (NAM) recognition based on iterative supervised adaptation. IEEE.Nara Institute of Science and Technology, Japan. 2003.

[45] M.Otani, S.Shimizu, andT.Hirahara, Vocal tract shapes of non-audible murmur production. The Acoustical Society of Japan. Sci. Technol, vol. 29, 195-198,2008.

[46] P. Heracleous, V. Tran, T.Nagai, andK. Shikano, Analysis and Recognition of NM Speech Using HMM Distances and Visual Information,IEEE transactions on Audio, speech, and language processing, vol. 18, no.6, 2010.

[47] D.Babani,T.Toda,H.Saruwatari, and K. Shikano, Acoustic model training for non-audible murmur recognition using transformed normal speech data, Nara Institute of Science and Technology, Japan. 2011.

[48] J.Freitas,A.Teixeira,M.Sales, andC.Bastos, Towards a Multimodal Silent Speech Interface for European Portuguese,Universidade de Aveiro, Portugal, 2011.

[49] V.Tran, G.Bailly, H.Loevenbruck andC. Jutten, "Improvement to a NAM captured whisper-to-speech system".Speech communication, vol. 52, issue 4, pp. 314-326, April 2010.

[50] P. Xiaomei, J. Hill, and G.Schalk, "Silent Communication: Toward Using Brain Signals"Pulse, IEEE, vol.3, no.1, pp.43-46, Jan. 2012.

[51] G.S.Meltzner,G. Colby, D.Yunbin, and J. T. Heaton, "Signal acquisition and processing techniques for sEMG based silent speech recognition", Engineering in Medicine and Biology Society,EMBC, 2011 Annual International Conference of the IEEE,pp.4848-4851, Aug. 30 2011-Sept. 3 2011.

[52] S. Ishii, T. Toda, H.Saruwatari,S.Sakti, and S. Nakamura,"Blind noise suppression for Non-Audible Murmur recognition with stereo signal processing",Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on , vol., no., pp.494-499, 11-15 Dec. 2011.